LUME - Módulo de estatísticas 2.0 para o Dspace 5.8

Cleusa Pavan[1]; Manuela Klanovicz Ferreira[2]; André Rolim Behr[3]; Caterina Marta Groposo Pavão[4]; Janise Silva Borges da Costa[5]

[1] Universidade Federal do Rio Grande do Sul. Email: cleusa.pavan@cpd.ufrgs.br
[2] Universidade Federal do Rio Grande do Sul. Email: manuelakf@cpd.ufrgs.br
[3] Universidade Federal do Rio Grande do Sul. Email: andre.behr@cpd.ufrgs.br
[4] Universidade Federal do Rio Grande do Sul. Email: caterina@cpd.ufrgs.br
[5] Universidade Federal do Rio Grande do Sul. Email: janise@cpd.ufrgs.br

Resumo

O monitoramento e a análise dos padrões de utilização dos repositórios institucionais permitem aos gestores avaliar e aprimorar a ferramenta. Os mantenedores de repositórios têm, dentre outras, a função primordial de aumentar a visibilidade de sua produção científica e, consequentemente, ampliar a leitura, a citação e a socialização do conhecimento. O trabalho descreve o Lume - Módulo de estatísticas 2.0, desenvolvido com base na versão disponível no DSpace 5.8, que permite a visualização gráfica dos dados de acessos e *downloads*, não disponível na versão original do DSpace. As estatísticas implementadas foram: a) número de *downloads* de todo o repositório, discriminados por comunidade, subcomunidade e coleção; b) os dez itens com maior número de *downloads* em cada subcomunidade de segundo nível; c) acesso e *downloads* agrupados por comunidade, subcomunidade, coleção ou item; d) por autor ou assunto, ambas permitindo o filtro de determinado período com visualização de gráficos por ano ou por mês, por top 10 países ou por todos os países de origem. Aborda, também, como adicionar o Módulo ao DSpace 5.8, a partir de um repositório público de código fonte.

Palavras-chave: Repositórios institucionais. DSpace 5.8. Estatísticas. Visibilidade.

Abstract

Monitoring and analyzing the usage patterns of institutional repositories allows managers to evaluate and improve the tool. Repository maintainers have, among others, the prime function of increasing the visibility of their scientific production. Consequently, increasing the reading, quotation, and socialization of knowledge. The paper describes Lume - Statistics Module 2.0, developed based on the version available in DSpace 5.8, which allows graphical visualization of access and download data, not available in the original version of DSpace. The statistics implemented were: a) number of downloads of the entire repository, broken down

by community, subcommunity, and collection; b) the ten items with the highest number of downloads in each second level subcommunity; c) access and downloads grouped by community, subcommunity, collection, or item; d) by author or subject, both allowing the filter of a defined period with graph visualization by year or by month, by top 10 countries or by all countries of origin. It also covers how to add the Module to DSpace 5.8, from a public repository of source code.

Keywords: Institutional repositories. DSpace 5.8. Statistics. Visibility.

Introdução

Os repositórios institucionais (RI) em acesso aberto, ao armazenarem e disponibilizarem a produção científica das instituições de ensino e pesquisa colaboram para aumentar a sua visibilidade e o seu uso. A visibilidade é um dos indicadores utilizados pelos sistemas de avaliação para determinar a posição de prestígio de um pesquisador, uma instituição ou um país. De acordo com Leite (2009), a contribuição principal dos RI está na reformulação e melhoria da comunicação científica ao adotar processos de gestão da informação científica.

Além desses aspectos, os RI fornecem dados relativos à sua utilização, podendo evidenciar o comportamento dos usuários em relação à ferramenta e ao conteúdo disponibilizado. Ao fazer *download* do texto, e não apenas visualizar os resultados de busca, o usuário manifesta interesse pelo documento recuperado. Por outro lado, estas informações também podem impulsionar os gestores dos RI a aprimorar e desenvolver novas funcionalidades.

O DSpace é a plataforma mais empregada pelos RI, com 43% do total de 4.140, conforme contabiliza o OpenDOAR (DIRECTORY OF OPEN ACCESS REPOSITORIES, 2019). A segunda posição é ocupada por EPrints com 13% e as demais plataformas são usadas por 3% ou menos dos repositórios.

O Lume, repositório digital da Universidade Federal do Rio Grande do Sul (UFRGS), foi desenvolvido utilizando o DSpace e implantado em 2008. Em maio de 2019, disponibilizava um pouco mais de 211.000 itens, abrangendo todas as áreas do conhecimento.

Desde 2011, dispunha de um módulo de estatísticas customizado com novas formas de visualização de acessos e *downloads*, onde as informações eram armazenadas e consultadas em uma tabela do banco de dados relacional, no caso o PostgreSQL. Conforme aumentava a quantidade de dados armazenados na tabela, a visualização passou a apresentar baixa *performance*. Mesmo com modificações que previam o pré-processamento de consultas, resumindo os dados antecipadamente, e a implantação de *cache*, não foi possível alcançar uma *performance* desejável.

A partir disso e, preparando a migração do Lume para o DSpace 5.8, em 2018, analisou-se a forma como os dados eram armazenados e consultados nesta versão. Constatou-se que o registro era feito no SOLR, uma plataforma de pesquisa e base de dados documental que permite consultas complexas aos seus

dados. Esta versão do DSpace disponibiliza, originalmente, três tipos de estatísticas: de uso, de busca e de fluxo de submissão, sem representação gráfica. As estatísticas podem ser visualizadas para todo repositório, por comunidade, subcomunidade, coleção ou item, e permitem verificar a quantidade de acessos dos últimos sete meses, por país e por cidade.

O Lume - Módulo de estatísticas 2.0, foi desenvolvido com o objetivo de gerar novas opções de agrupamento e visualização dos dados, sob a forma de tabelas e gráficos, mediante consulta no registro *default* de estatísticas do DSpace 5.8, facilitando a leitura. Além disso, visa sanar o problema de desempenho do módulo implantado em 2011.

Implementação técnica do Módulo

Após a análise e a realização de testes da versão original do DSpace 5.8, modificou-se o antigo módulo de estatísticas do Lume para que consultasse a interface SOLR, disponível no DSpace 5.8, ao invés do banco de dados PostgreSQL. Os primeiros resultados não alcançaram a *performance* esperada, mesmo tendo a consulta mais rápida que no anterior. Realizados novos estudos e testes, foram trabalhados dois pontos:

a) a consulta feita por período aos dados no SOLR estava muito demorada, então decidiu-se criar índices resumidos por ano e por mês;
b) a consulta por facetas, disponível no SOLR, foi otimizada de 10 segundos para 1.5 quando utilizado o parâmetro de consulta do SOLR facet.query, ao invés do parâmetro facet.field, conforme documentado em pergunta no fórum StackOverflow[1].

Ao atingir a *performance* desejada, foi feito o *release* com as seguintes modificações necessárias para implementar o Lume - Módulo de estatísticas 2.0:

a) alteração do <dspace-source>/dspace/solr/statistics/schema.xml> com adição de dois campos resumidos de data por ano e por mês, e reindexação do *core* de estatísticas do SOLR (Figura 1);
b) inclusão e alteração de arquivos fonte em:
   - JAVA: para a consulta das estatísticas do SOLR e a geração de DRI foi incluída uma classe JAVA para cada tipo de visualização das estatísticas;
   - JAVASCRIPT: para a geração dos gráficos foi utilizada a biblioteca chart.min.js[2] e também foi criado o arquivo chartEstatisticas.js;
   - CSS: foram incluídas regras para o posicionamento das estruturas nas páginas de visualização.
c) modificação no arquivo <theme>/xsl/core/navigation.xsl> de menu lateral para incluir o *link* para as estatísticas nas respectivas páginas;

[1] https://stackoverflow.com/questions/51247635/debug-a-solr-query-with-facet-field-shows-facet-result-over-more-documents-than
[2] http://chartjs.org/

```
<!--Adicionando para conter a data apenas com ano e mes-->
<fieldType name="keywordTimeFilterAnomes" class="solr.TextField" sortMissingLast="true" omitNorms="true">
    <analyzer>
        <charFilter class="solr.PatternReplaceCharFilterFactory"
         pattern="(^\d{4}-\d{2})(.+)" replacement="$1"/>
        <!--Treats the entire field as a single token, regardless of its content-->
        <tokenizer class="solr.KeywordTokenizerFactory"/>

        <!--<filter class="solr.LowerCaseFilterFactory" />-->
        <filter class="solr.TrimFilterFactory" />
    </analyzer>
</fieldType>

<!--Adicionando para conter a data apenas com ano e mes-->
<fieldType name="keywordTimeFilterAno" class="solr.TextField" sortMissingLast="true" omitNorms="true">
    <analyzer>
        <charFilter class="solr.PatternReplaceCharFilterFactory"
         pattern="(^\d{4})(.+)" replacement="$1"/>
        <!--Treats the entire field as a single token, regardless of its content-->
        <tokenizer class="solr.KeywordTokenizerFactory"/>

        <!--<filter class="solr.LowerCaseFilterFactory" />-->
        <filter class="solr.TrimFilterFactory" />
    </analyzer>
</fieldType>

[...]

<!--adicionado para diminuir o tempo-->
<field name="time_anomes" type="keywordTimeFilterAnomes" />
<field name="time_ano" type="keywordTimeFilterAno" />
<copyField source="time" dest="time_anomes" />
<copyField source="time" dest="time_ano" />
```

Figura 1: Alteração do arquivo <dspace-source>/dspace/solr/statistics/schema.xml>

d) modificação do sitemap.xmap para suportar os *links* para as estatísticas (Figura 2).

O Módulo aproveita os registros *default* de estatísticas coletadas pelo DSpace 5.8, através de consulta à interface SOLR, o que minimiza o número de alterações necessárias para a sua implantação e permite que as estatísticas originais permaneçam ativas. Além disso, permite que, no futuro, sejam pensadas novas formas de visualização para as estatísticas de busca e fluxo de submissão, que também são registradas por *default* no SOLR.

Ao registrar as estatísticas *default*, o DSpace 5.8 identifica quais são resultantes das buscas automáticas feitas no Repositório pelos robôs dos motores de busca, a fim de indexar o seu conteúdo. Essas estatísticas ficam marcadas com “isBot:true” e não são contabilizadas na exibição para o usuário final.

O *download* do Módulo e a identificação das modificações necessárias estão disponíveis no GitHub.

```
<map:sitemap xmlns:map="http://apache.org/cocoon/sitemap/1.0">
    <map:components>
        <map:transformers>
            <map:transformer name="BrowseStats" src="org.dspace.app.xmlui.aspect.statistics.BrowseStats"/>
            <map:transformer name="Downloads" src="org.dspace.app.xmlui.aspect.statistics.Downloads"/>
            <map:transformer name="Stats" src="org.dspace.app.xmlui.aspect.statistics.Stats"/>
            <map:transformer name="TopTen" src="org.dspace.app.xmlui.aspect.statistics.TopTen"/>
        </map:transformers>
    </map:components>
    <map:pipelines>
        <map:pipeline>
            <map:generate/>

            <map:match pattern="stats/downloads">
                <map:transform type="Downloads"/>
            </map:match>

            <map:match pattern="stats/topten">
                <map:transform type="TopTen"/>
            </map:match>

            <map:match pattern="browsestats">
                <map:transform type="BrowseStats"/>
            </map:match>

            <map:match pattern="handle/*/*/browsestats">
                <map:transform type="BrowseStats"/>
            </map:match>

            <map:match pattern="handle/*/*/stats">
                <map:transform type="Stats"/>
            </map:match>

            [...]
            <!-- Not a URL we care about, so just pass it on. -->
            <map:serialize type="xml"/>

        </map:pipeline>
    </map:pipelines>
</map:sitemap>
```

Figura 2: Alteração do arquivo sitemap.xmap

As estatísticas de todos os níveis do Lume estão disponíveis para o público sem a necessidade de *login*. Podem ser consultadas por comunidade, subcomunidade, coleção ou item e são identificadas por meio de ícone próprio:

Exibição das estatísticas

As estatísticas são apresentadas a partir da aba Sobre, opção Estatísticas gerais ou vinculadas a cada nível do Repositório, autor e assunto.

Para fins de ilustração dos exemplos apresentados nas alíneas a e b, adotou-se a comunidade Produção científica.

O módulo disponibiliza as seguintes estatísticas:

a) número de *downloads*

O total geral de *downloads*, sob a forma de lista, é exibido numa única página, bem como o número de *downloads* por comunidade, subcomunidade e coleção (Figura 3).

É possível ter um panorama de *downloads,* de acordo com a organização das comunidades adotadas pelo Lume, que abrangem órgãos específicos da UFRGS, programas de pós-graduação, áreas do conhecimento e tipos de materiais. Este número, combinado com outros dados que podem ser extraídos pelo administrador da plataforma DSpace, a partir de tabelas de logs, viabiliza análises e estudos métricos, contribuindo para a gestão da produção institucional.

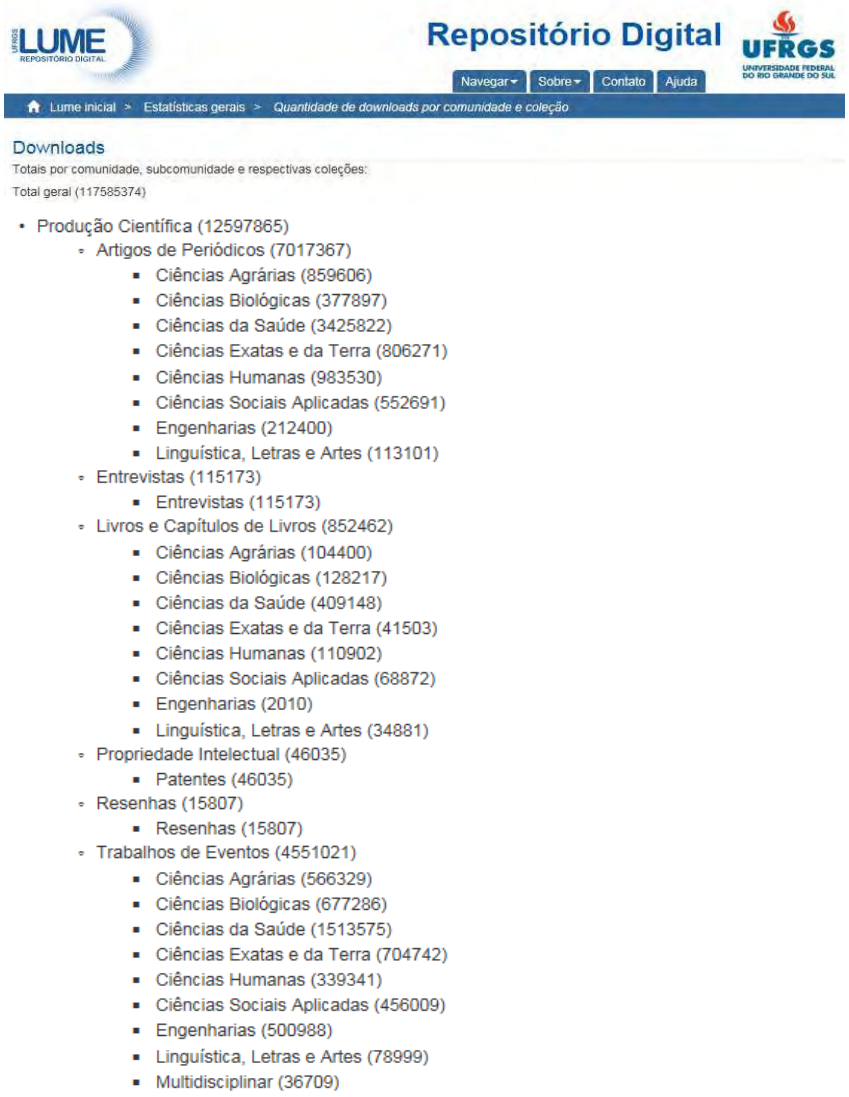


Figura 3: Número de *downloads*

b) itens com maior número de *downloads* (top 10)

A Figura 4 lista o título dos dez itens com maior número de *downloads* no segundo nível da hierarquia do Lume e inclui este dado entre colchetes.

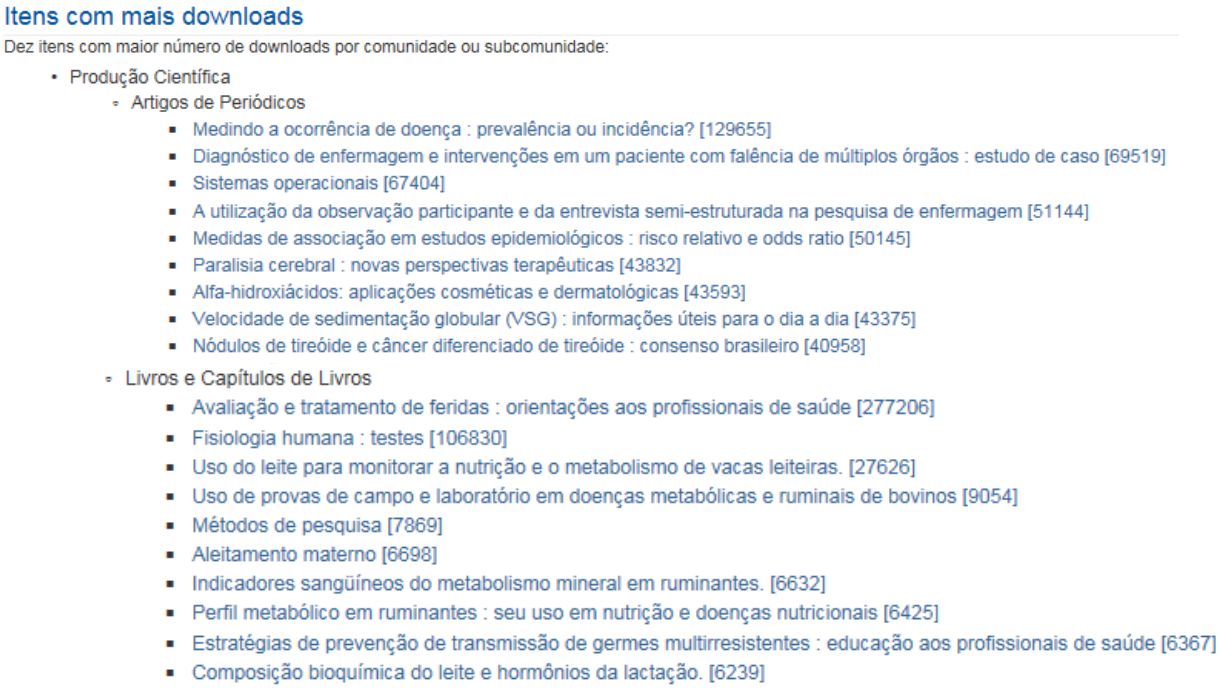


Figura 4: Dez itens com maior número de *downloads*

c) número de acessos e *downloads*

Está disponível por meio de ícone próprio na página principal da comunidade, subcomunidade, coleção ou item.

Os dados são apresentados sob a forma de tabelas e gráficos, distribuídos anualmente (Figura 5). É possível visualizar, também, a distribuição mensal dos dados e restringir o período temporal, utilizando filtro de ano e mês. Ao final da página são relacionados, por *default*, os dez países com maior número de acessos e de *downloads*, com a opção de consulta à lista completa de países.

Estatísticas

Início da coleta: Jan. 2008

Teses e Dissertações defendidas na UFRGS

Estatísticas por ano por mês

Filtrar por data

	Ano	Mês
Data inicial:	Todos	Todos
Data final:	Todos	Todos

Enviar

Downloads por país

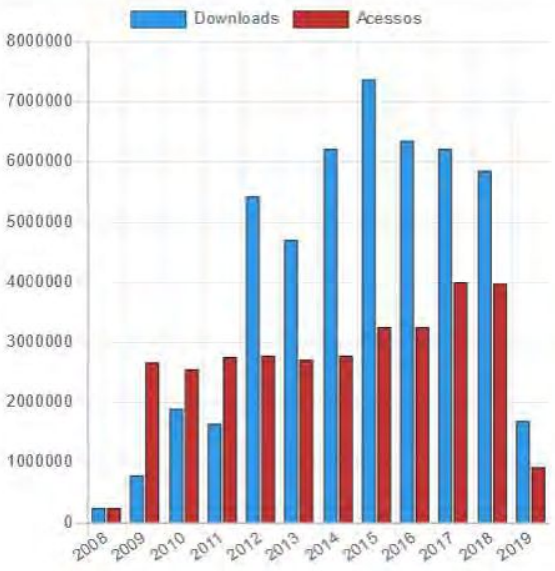
País	Downloads
Brasil	34047534
Alemanha	5812736
Estados Unidos	2458414
China	1648995
Portugal	1385198
Moçambique	471729
UFRGS*	396855
? Outros	285105
França	271665
Angola	263901

Acessos por país

País	Acessos
Brasil	23592250
Estados Unidos	3504256
China	1245240
Portugal	919825
Alemanha	602283
UFRGS*	522717
França	502035
? Outros	151591
Japão	148849
Moçambique	141890

Estatísticas por ano

Ano	Downloads	Acessos
2008	231915	233535
2009	784272	2668554
2010	1883464	2543790
2011	1633944	2745703
2012	5421370	2773181
2013	4701780	2715252
2014	6217634	2773179
2015	7361810	3243374
2016	6351957	3249648
2017	6219448	4000839
2018	5845246	3982978
2019	1683438	928675
Total	48336278	31858708



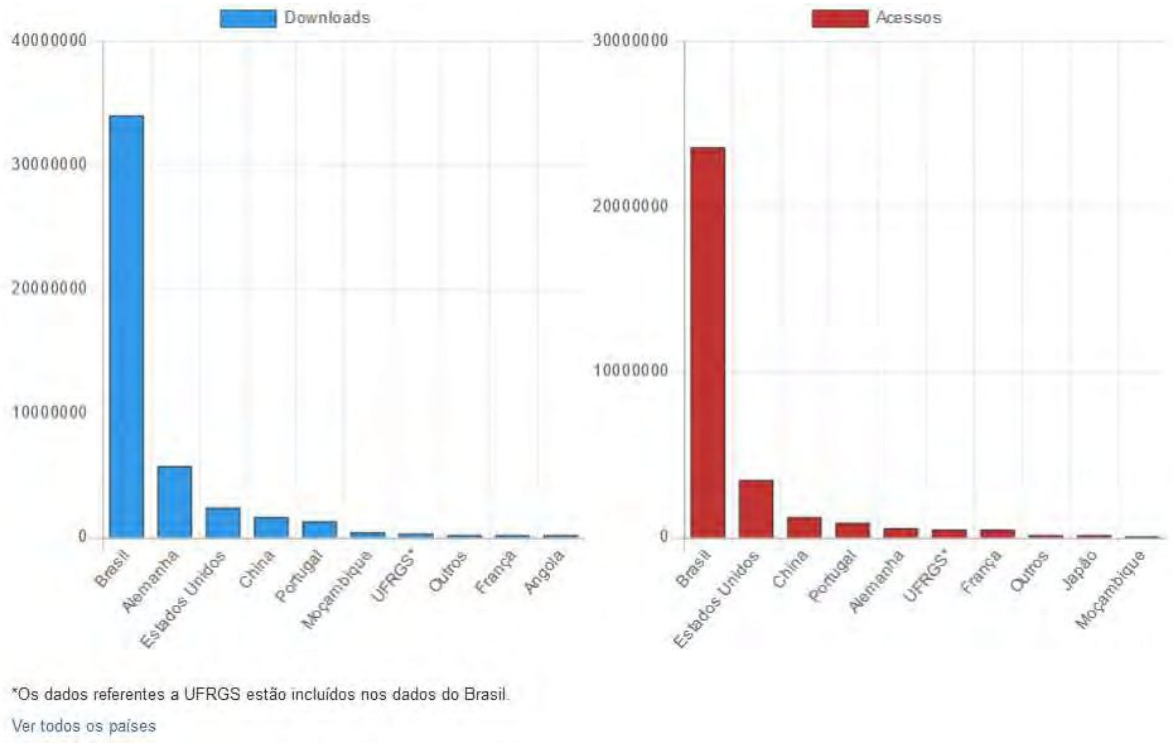


Figura 5: Número de acessos e *downloads* na subcomunidade de Teses e Dissertações defendidas na UFRGS

d) itens de um determinado autor ou assunto

Os dados são gerados a partir dos itens recuperados nos índices de autor e de assunto. Após a realização da busca, o link do ícone próprio das estatísticas fica no topo da lista de resultados (Figura 6).

Ao clicar no ícone, são listados, inicialmente, os cinco itens com maior e os com menor número de *downloads* do total de determinado autor depositados no Lume ou de assunto, seguidos do número de *downloads* e da data de entrada do item no repositório, conforme ilustra a Figura 7.

Na sequência, é exibida a distribuição mensal dos dados, com possibilidade de restrição do período temporal, e são apresentados, por *default*, os dez países com maior número de acessos e de *downloads*, além de *link* para a lista completa de países.

Figura 6: Itens para um autor

Figura 7: Itens com maior e menor número de *downloads* para um autor

Considerações finais

O Lume - Módulo de estatísticas 2.0, agrega valor aos repositórios que utilizam o DSpace 5.8, ao possibilitar a recuperação e visualização gráfica de dados de acessos e de *downloads* nos vários níveis hierárquicos e, sobretudo, aos itens relacionados a um mesmo autor ou assunto.

Com a consulta aos dados já registrados pelo DSpace 5.8 no SOLR, foi possível melhorar o desempenho na recuperação e geração das estatísticas, bem como manter aquelas registradas originalmente pelo DSpace. Assim sendo, caso sejam criadas novas possibilidades de estatísticas no futuro, a integração será facilitada, pois o Módulo consulta as estatísticas padrão armazenadas pela ferramenta.

Referências

DIRECTORY OF OPEN ACCESS REPOSITORIES. OpenDOAR Statistics. 2019. Acesso em: 9 maio 2019. Disponível em: <http://v2.sherpa.ac.uk/view/repository_visualisations/1.html>.

LEITE, F. C. L. **Como gerenciar e ampliar a visibilidade da informação científica brasileira:** repositórios institucionais de acesso aberto. Brasília: IBICT, 2009. 124 p. Disponível em: <http:/livroaberto.ibict.br/handle/1/775>. Acesso em: 18 maio 2019.